DOCUMENT RESUME

ED 272 522 TM 860 427

AUTHOR Thomas, Gregory P.
TITLE Information Needs within a Multi-District Environment.
PUB DATE 17 Apr 86
NOTE 8p.; Paper presented at the symposium, "Taming the Rasch Tiger: Using Item Response Theory in Practical Educational Measurement," at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 17-20, 1986). For other symposium papers, see TM 860 424-428.
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Adaptive Testing; *Criterion Referenced Tests; Data Collection; Difficulty Level; Information Needs; Item Banks; *Latent Trait Theory; Mathematical Models; Measurement Techniques; *Standardized Tests; *Test Construction; Test Items; Test Theory
IDENTIFIERS Rasch Model

ABSTRACT

This paper argues that no single measurement strategy serves all purposes and that applying methods and techniques which allow a variety of data elements to be retrieved and juxtaposed may be an investment in the future. Item response theory, Rasch model, and latent trait theory are all approaches to a single conceptual topic. An abbreviated look at each technique explains its particular application. Standardized achievement tests which are prepared by publishing companies are criticized because they evaluate broad areas of curriculum without measuring the particular curriculum of a school district. With the availability of item banks, many individuals develop local criterion referenced tests. This process is time consuming and produces results which over time may not be desirable. As more data becomes available, the emphasis should shift from a particular measurement technique to the development of schemes to examine data drawn from various sources, reflecting a variety of perspectives. Facing the limitations of any testing technique, recognizing the limitations of data collection processes, and applying additional solutions to information problems promises new frontiers to the measurement community. (JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED272522

INFORMATION NEEDS WITHIN A MULTI-DISTRICT ENVIRONMENT

Presented at the

National Council for Measurement in Education
San Francisco, California
April 17, 1986

by

Gregory P. Thomas
Washington County Education Service District

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. P. Thomas

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.
☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 860427

INFORMATION NEEDS WITHIN A MULTI-DISTRICT ENVIRONMENT

Presented at the

National Council for Measurement in Education
San Francisco, California
April 17, 1986

Gregory P. Thomas
Washington County Education Service District
17705 Springville Road
Portland, Oregon 97229

The Northwest corner of the United States is populated by a large group of professionals who would have one believe that creating a curriculum matched, difficulty indexed, individually oriented test was relatively easy. They point out the decade-plus history of item bank development, and the availability of inexpensive computer hardware as well as software.

An equally vocal group of professionals will point out the need to meet state mandated testing requirements, amortize investments in nationally normed test instruments, stay compatible with long data histories of student performance, and the importance of statistics and reported results which can be understood and explained to diverse audiences. Add to this mix the "new" thrust on school effectiveness and "profiling" and the seemingly simple issue of data and information becomes very complex.

I intend to argue in this paper that the simplistic approaches which many people urge will not produce information of utility in the management of the complex environment known as a school district. I will also explore the single data point notions prevalent today and suggest an alternative or two which might be used as one attempts to develop usable data for management (teachers also manage under my notion of management, not just administrators).

ITEM RESPONSE THEORY (IRT) - RASCH MODEL - LATENT TRAIT

Although each of these terms is definable and much writing has occurred (Birnbaum, 1968), for convenience, I will deal with them as a single conceptual topic. The basic test development procedure used in these approaches begins with the development of test items by a team of "trained individuals." These individuals write items. The items are then put into a test format and data are collected from an examinee sample of convenience. One of a number of statistical treatments is then applied to identify items which do not meet the constraints of the particular statistical model being employed. Items are then either discarded or re-written.

This process is straight forward and familiar to all who have developed tests. The difference is that in using item response theory (IRT) the items are assigned a "difficulty index" based on their relationship to the other items in that testing session. The index is to serve from this point on as a principal key to item identification and use in future test development.

It is worth taking an abbreviated look at some of the techniques which are used in this process as well as the assumptions which must be adopted. A brief look at the Rasch model, one widely used in our area, will illustrate

many of the assumptions. This particular model statistically contrasts curriculum difficulty of test items with the probability of an individual of certain ability answering the item correctly. "Good" items will show high discrimination if they are to be retained. In short, if a person is of "high ability", the item should be correctly answered and, at the same time, incorrectly answered by a person of low ability. Whereas the Rasch model represents a single parameter model, much has also occurred as a result of the introduction of multiple parameter models (Lord and Novick, 1968). Multiple parameter models are akin to the exercise of adding variables to a multiple regression so that the error term is further and further refined in an attempt to enhance accuracy of the measure.

The samples of individuals used to generate item data are most typically "samples of convenience." Classrooms of students and, in some cases, school districts have been utilized as test sites for new items. The size of needed samples is accepted as 200 based on work done by the Portland Public Schools Evaluation Department (occasional paper, P.P.S.). Because of the pervasive belief that item difficulty does not vary across samples of differing ability, more traditional sampling approaches are not typically used.

However, holding the item pool constant but varying the examinees produces a somewhat different result. Reviewing student performance data across a two year span revealed that item difficulty will vary between administrations of a constant set of indexed items (occasional paper, MESD). That is, the difficulty of some items remained stable across different populations of students while some item difficulties changed. This suggests that we should expect difficulty drift over time and must constantly re-analyze and link back to a baseline set of data if we want to be able to use a stable scale (Rudner, 1983). This re-analysis is not discussed nor of concern to most users of indexed item banks. It is suspected that this phenomenon is not commonly understood nor is there a process which could be put in place to accommodate test developers which would be cost effective and simple to use.

In sum, item response theory proposes that item difficulty can be defined at a point in time, that difficulty is constant across populations and will remain constant over time. Adapting tests to an individual's ability is a matter of selecting items which match that individual's capabilities and given the tools available constructing tests is relatively painless.

Re-analysis of test data where the items are held constant but the examinee population is varied suggests that item difficulty is not as stable as proported for all items on a given instrument and in all contexts. This item drift has been noted both in a mathematical as well as an applied sense. In order to account for this item drift, a statistical link back to baseline data must be performed. Linking back to baseline with every test administration may represent a level of complexity which is well beyond the casual test developer's capabilities. Ironically, it was this group of individuals that these techniques were supposed to directly support. From an operational standpoint this suggests a commitment of staff, time and resources are necessary to maintain and operate a testing program based on IRT that may be beyond the grasp of most.

STANDARDIZED ACHIEVEMENT TESTS

Instruments which are prepared by publishing companies tend to follow statistical approaches which are more traditional. These instruments are often criticized for not meeting the exact curricular approaches used within a particular school district. Comments such as "...tests of this ilk measure broad areas but not our particular curriculum..." are common place. These statements indicate that more precise measurement is somehow needed and desirable.

It is certainly factual that publishing company tests must cover a very broad spectrum of curriculum. It is also the case that this spectrum is probably contained in one form or another in virtually every classroom. This suggests that there are also large segments of the curriculum which are not tested with these instruments. If Bracey (1986) is correct in his analysis, and there is no better than a 50% match between standardized tests and curriculum, placing heavy reliance on point-in-time assessments will produce minimal usable information. Perhaps the debate regarding match and precision of measurement, IRT vs. standardized tests, is of far less concern than long term need for additional categories of data.

In sum, publisher tests measure everyone's curriculum and at the same time no ones specific adaptations. They are broad in scope and most likely cover the curriculum elements most individuals define as "basic skills." They produce a data set which will aggregate from an individual item response to a district wide summary of performance. Perhaps their most significant difference in contrast to IRT based tests is the availability of a contrast population as opposed to a difficulty index. The question for the user of these instruments must therefore be one of minor advantages. Selecting the CTBS U/V, the Portland Area Levels Test, CAT or SBS should probably be based more on the quality of reports and the degree of test/curriculum match than the item analytic and scale theory used to generate the test. Even more important is the relative contributions of these data within a larger information framework.

CRITERION REFERENCED - LOCALLY DEVELOPED

With the availability of item banks, many individuals have chosen the path of developing local tests. These individuals typically will organize a committee of teachers whose charge is first to define the curriculum which is in place. Using item banks, these individuals are asked to select items which match these objectives. Most frequently coverage of all goals is not available through these item banks. In this case these individuals are asked to develop specific items to cover the materials. Sometimes training on item writing is provided, some times not. Once a test has been constructed, typed, proofed and reproduced the committee disbands believing their job complete.

In some instances a pilot administration of the developed instrument is possible and practical. Most often the results of these pilot administrations reveal items which do not discriminate, based on IRT or more traditional item analysis techniques. This cycle in some test development efforts is repeated until such time that statistical analysis indicates that all included items do discriminate. Experience in these projects indicates that a minimum of three to four months is needed to complete this cycle. This is a considerably different result than drawing items from a pool and having a test ready to be used.

An interesting question comes to mind in thinking through the potential effects of deleting items. If we repeatedly delete items from tests which are not meeting mathematical model expectations, over the course of time we may be producing a test which is free of curriculum effects. Perhaps following this approach in an attempt to produce a "tight" metric serves the purpose of retaining items which are in fact not impacted by those modifications to curricula which we continually strive to make. This may serve to insure that the resultant instrument is no more or less sensitive to curriculum than one published by a publishing house.

Individuals who participate in these experiences do appreciate not only the difficulty of test development but also that of defining a measurable unit of the curriculum. Tests are refined and sharpened to a level of specificity far removed from that which guided the initial project. Once these instruments are constructed they also tend to remain in place for extended periods of time. The idea of bringing a committee back together to redo a test to reflect a change to the adopted series is not frequently discussed. Using batteries of ill maintained item banks and tests leaves one wondering about the magnitude of problems which are being created at the local district level.

In sum, producing CRT's from existing item banks is not a simple process. It does consume enormous energy and time and produces a result which over time may not be desirable. Integrating these instruments into a comprehensive testing program is often confusing in terms of defining a nich which seems most appropriate. Given the drift of item difficulty as well as the difficulty of maintaining tests, perhaps infrequent test developers are more appropriately advised to refrain from developing local instruments based on difficulty indexed items. At least item banking should not be embraced as a panacea for more precision or better curricular fit in educational measurement.

INFORMATION SYSTEMS

What I have pointed out above is that item banks and associated metric approaches, in similar fashion to other metric approaches, have shortcomings. If used in a supervised fashion they are capable of producing usable data. Instruments produced by publishing companies also produce a level of usable data. The instruments developed at a local level also have a place in the data scheme of things. It is important to not become so tied to a particular methodology, however, that data options begin to be dictated by the method. Another way of saying this is the method becomes the objective.

From a practitioner's perspective the proported accuracy and precision of measurement gained from a particular approach loses itself in the realities of daily school operation. Data collected at points in time are valuable as describing that point in time. The particular measurement approach is not the key issue. The utility of any data point, is from a broader perspective, key to the energy dedicated to the collection of those data. While each field needs its champions, measurement being no exception, perhaps some tempering of technique with operational reality needs to occur.

The thrust on "school effectiveness" has begun to focus our attention on the notion of multiple data points collected over extended periods of time. The scores which we generate from a particular test instrument take a place juxtaposed with grade histories, attendance patterns, changes in family status and so on. This is beginning to suggest that as more data become available to us, the emphasis should begin to shift from a particular measurement technique to

the development of schemes to examine data drawn from various sources reflecting a variety of perspectives.

Focusing energy on the organization and recombination of existing data appears to, at minimum, have promise for a large return in usable information. Facing the limitations of any testing technique or program squarely seems most appropriate. Recognizing limitations of data collection processes and moving forward with additional solutions to information problems promises many new frontiers for the measurement community.

SUMMARY

I have attempted to point out that no single measurement strategy will serve all purposes. The idea that we can examine an item at a single point in time with a sample of convenience and have that item remain constant in perpetuity seems to not be the case. With some individuals believing that accurate measurement of no more than 50% of the curriculum will occur, no matter what the instrument, then expending more and more energy on developing a "more accurate metric" seems to be time not well spent.

Investing in methods and techniques which allow a variety of data elements to be retrieved and juxtaposed may be a far better investment in the future. If we are able to recombine diverse data points to address a specific question regarding a trend, a student, a building and so on, we may be more closely responding to the goal of providing a better managed educational environment for the students whom we serve.

REFERENCES

Birnbaum, A., Some Latent Trait Models and Their Use in Inferring an Examinee's Abilities. In Lord, F.M. & Novick, M.R., Statistical Theory of Mental Test Scores. Addison-Wesley, 1968.

Bracey, G.W., Mismatch of Testing and Instruction is Pervasive. Phi Delta Kappan, March, 1986.

Lord, F.M., & Novick, M.R., Statistical Theory of Mental Test Scores. Addison-Wesley, 1968.

Occasional Papers, Multnomah Education Service District, Measurement and Evaluation Department.

Occasional Papers, Portland Public Schools, Departments of Research and Evaluation.

Rudner, L.M., A Closer Look At Latent Trait Parameter Invariance. Educational and Psychological Measurement, 43, 1983.